



WebCaricature: a benchmark for caricature face recognition

Jing **Huo**^a, Wenbin **Li**^a, Yinghuan **Shi**^a, Yang **Gao**^{a,**}, Hujun **Yin**^b

^aState Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210046, China.

^bElectrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, UK.

ABSTRACT

Caricatures are facial drawings by artists with exaggeration on certain facial parts. The exaggerations are often beyond realism and yet the caricatures are still recognizable by humans. With the advent of deep learning, recognition performances by computers on real-world faces has become comparable to human performance even under unconstrained situations. However, there is still a gap in caricature recognition performance between computer and human. This is mainly due to the lack of publicly available caricature datasets of large scale. To facilitate the research in caricature recognition, a new caricature dataset is built. All the caricature images and face images were collected from the web. Compared with two existing datasets, this dataset is of larger size and has various artistic styles. We also offer evaluation protocols and present baseline performances on the dataset. Specifically, four evaluation protocols are provided: restricted and unrestricted caricature verifications, caricature to photo and photo to caricature face identifications. Based on the evaluation protocols, three face alignment methods together with five kinds of features and nine subspace and metric learning algorithms have been applied to provide the baseline performances on this dataset. Main conclusion is that there is still a space for improvement in caricature face recognition.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In the past few years, face recognition performance, even in unconstrained environments, has improved rapidly. For instance, recognition accuracy on the LFW dataset has reached 99%, which even outperforms most humans [1]. However, there seems still a significant gap in caricature recognition performance between computer and human. Caricatures are facial drawings by artists with exaggerations of certain facial parts. The exaggerations are often beyond realism and yet the caricatures are effortlessly recognizable by humans. In some cases, people find that caricatures with exaggerations are easier to recognise compared to real face photos. However, this is not the case for computers. Caricatures have long fascinated psychologists, neuroscientists and now computer scientists and engineers in their seemingly and grossly distorted views of veridical faces but still possessing distinctively recognizable features of the subjects [2], [3], [4]. Studying caricatures can offer insights

to how face recognition is robustly performed. The study of caricature face recognition can also lead to better understanding of human perception of faces. With this objective, we have collected and set up a dataset of caricatures to promote and facilitate the study of caricature face recognition. Besides, we have provided face landmark annotations on this dataset, as well as evaluation protocols and baseline performance for comparison.

Currently, there are only two public available caricature datasets [5], [6]. The first set [5] includes pairs of caricature and photo of 196 subjects. There were two sources in this dataset. The first was from commissioning artists to draw caricatures (from photos) and resulting in 89 caricatures. The second was via the Google image search and 107 image pairs were collected in this way. The second dataset [6] includes caricature and photo pairs of 200 people. All the caricatures of this dataset are hand drawn black and white caricatures and of a single artistic style. Therefore, it is obvious that the existing caricature datasets are limited in subject size and artistic style. To motivate the better study of caricature face recognition, we have built a dataset of larger size with 6042 caricatures and 5974 photographs in total. These caricatures and photographs are from 252 persons. While building this dataset, the main focus

^{**}Corresponding author.

e-mail: gaoy@nju.edu.cn (Yang Gao)



Fig. 1. Examples of photos and caricatures. First column are photos and next six columns are corresponding caricatures. The caricatures are of various artistic styles, exhibiting large intra-personal variations.

was to cover as many caricatures of a person as possible. With each subject having caricatures of various artistic styles, this dataset is more suited for studying caricature face recognition compared to the existing two. Examples of photos and caricatures of the new dataset are shown in Fig. 1.

Main contributions of this paper are as follows:

A new caricature dataset of 252 persons is built. For each person, there are average 33 photos and 57 caricatures from various sources of the Internet. For each image in the dataset, we also provide 17 labeled facial landmarks. This is the largest caricature dataset available so far. For each person, we collected as many caricatures as possible to include as many artists as possible to have the variability in caricature style and exaggeration. To find what is more intrinsic for recognizing caricatures and related photos, we thus tried to cover more intra-personal variations than more subjects. As all the caricature images are collected from the web, the caricatures are of diverse artistic styles. Besides, the illumination conditions, pose, expressions, occlusions and age gaps in the caricature and photo modalities were not controlled. All these factors make the caricature recognition problem realistic and generic.

With this dataset, we also provide four basic experimental protocols. They are, restricted and unrestricted caricature verifications, caricature to photo and photo to caricature identifications, offering a way for fair comparison of any future work. Besides, we present a new framework for caricature face recognition, including caricature face detection, facial landmark detection, caricature face alignment, feature extraction and feature matching. Challenges for each of these steps are discussed.

With the framework and evaluation protocols, baseline performances are also provided. Specifically, we verified from the following aspects. For caricature and photo face alignment, we tested three methods. For feature extraction, four hand-crafted features and one deep learning based feature are tested. With respect to feature matching, nine subspace and metric learning methods are applied for comparison, including both single-modal and cross-modal based methods. A conclusion is that deep learning based feature performs the best. However their

performance can be further improved with subspace or metric learning. Besides, there is still a space for improvement, even for deep learning based methods.

2. Related Work

2.1. Caricature Face Recognition

Currently, there is limited work on caricature face recognition. The work of [5] proposed to use qualitative features to solve the problem. The qualitative features were labeled by human via Amazon’s Mechanical Turk service, and were then combined with logistic regression (LR), multiple kernel learning (MKL), and support vector machines (SVM) to calculate the similarity of a caricature and a photo. A dataset was released in this work, where there are 196 pairs of caricatures and photos. Authors in [7] proposed to learn a facial attribute model. These facial attribute features were then combined with low-level features for recognition using canonical correlation analysis (CCA). An attribute dataset for the dataset of [5] is released, in which each photo and caricature were labeled with 73 attribute features. Abaci and Akgul [6] proposed a method to extract facial attribute features for photos. For caricatures, the attribute features were manually labeled. Then the weights of these attribute features were learned by a genetic algorithm (GA) or LR for recognition. In this paper, the authors released a dataset of 200 pairs of caricatures and photos. All the caricatures of this dataset are line drawing caricatures. Crowley et al. [8] proposed to use fisher vector (FV) and convolutional neural network (CNN) representations combined with discriminative dimensionality reduction (DDR) or SVM classifiers to retrieve similar paintings given a query photo. The paintings studied in this work are of many painting media (oil, ink, watercolor) and various styles (caricature, pop art, minimalist). Two datasets have been built but are not publicly available. The first dataset is from the website DeviantArt. The second dataset is from National Portrait Gallery (NPG). A summary of these existing and related work is given in Table 1.

From the above, it is evident that currently there is little work on caricature face recognition, despite its imperative value to face perception and recognition is commonly recognised. This is largely due to limited datasets available and limited scale of the current caricature datasets. Therefore, we have built a new dataset to facilitate the study on caricature recognition.

2.2. Heterogeneous Face Recognition

Caricature face recognition is a special case of heterogeneous face recognition where the two matching face images are of different modalities. According to the way to remove modality variations, there are generally three kinds of approaches to the recognition problem.

The first is image synthesis based methods. The main idea is to map images of one modality into another by image synthesis [9], [10]. The second is modality invariant feature extraction based methods [11], [12]. Designed feature descriptors or deep learning methods are used to extract face features that are invariant to modality change. The last is common subspace based methods, which try to find a common subspace for face

Table 1. Summary of existing work on caricature recognition.

	Dataset Information			Features	Learning Method
	Number of subjects	Images	Available		
Klare et al. [5]	196 subjects	392 images	Yes	Attribute feature	LR/MKL/SVM
Ouyang et al. [7]	-	-	-	Attribute feature/ low-level feature	CCA
Abaci and Akgul [6]	200 subjects	400 images	Yes	Attribute feature	GA/LR
Crowley et al. [8]	Dev: 1,088 subjects NPG: 188 subjects Train: 496 subjects	Dev: 8,528 images NPG: 3,128 images Train: 257,000 images	No	FV/CNN feature	DDR/SVM
This paper	252 subjects	12,016 images	Yes	Gray/LBP/Gabor/ SIFT/CNN feature	9 Subspace learning and metric learning methods

features of two modalities. Then the recognition is performed in the common subspace, as the projections to the common subspace are sought to remove modality variations. Related work includes common discriminant feature extraction (CDFE) [13], coupled spectral regression (CSR) [14] and so on.

Different from the existing heterogeneous face recognition applications, caricatures have exaggerated face parts, making caricatures beyond realism. Such exaggerations do not seem to cause difficulties to humans in recognition. However, this is not the case for computers. Thus it is important to study caricatures in terms of revealing features that are more intrinsic for face recognition.

3. Dataset Collection

3.1. Collection Process

For collecting both face images and caricatures, we first drew up a list of names of celebrities. For each of these people, we manually searched for their caricatures and photos and saved these images. All the photos were searched using the Google image search. For caricatures, they are mainly from Google image search and Pinterest. After all the images were downloaded, a program was written to remove duplicated caricatures and photos. This was done by extracting features of caricatures and faces of photos, two features with a large similarities were picked out. We then manually checked if they came from duplicated images. This resulted in a dataset of 252 people with a total of 6042 caricatures and 5974 photographs. For each subject, its number of caricatures ranges from 1-114 and number of photos from 7-59.

3.2. Labeling Information

For each caricature, 17 landmarks were manually labeled as shown in Fig. 2. The first four landmarks are face contours. The next four are eye brows. Landmarks 9-12 are eye corners. The 13th landmark is the nose tip. The 14th-17th are mouth contours.

The labeling procedure for photo was different. For each photo, we firstly used the facial landmark detection software provided by Face++ [15]. This software could locate 83 face landmarks, from which we used the 2nd-17th face landmarks that correspond to our manual labeled landmarks for caricatures. The first landmark was labeled manually for each photo.

**Fig. 2. Example of labeled 17 facial landmarks.**

Note that there were some photos the software did not label well or failed to locate the landmarks. For those cases, we manually corrected the labeling or manually labeled them with the same scheme as for caricatures. After the above two processes, 17 landmarks for each caricature and photo were obtained.

4. Recognition Framework and Challenges

As reviewed in Section 2, most existing work on caricature recognition tries to extract descriptive attributes or qualitative features for recognition. In this paper, we illustrate how to solve the recognition problem without using attributes, as they are subjective and require extensive manual labeling. The recognition framework is shown in Fig. 3. The process is similar to the traditional face recognition process. However, there are challenges at almost each step in this case.

4.1. Caricature Face Detection and Landmark Localization

The first step of the recognition process is to perform caricature face detection and landmark localization. For the detection task, as caricatures can vary greatly in artistic styles and facial appearances. It is more difficult to find the detection patterns of caricatures compared with real faces in photos. And for the task of localizing landmarks, the shapes of caricatures and the positions of eyes, noses and mouths can be exaggerated and sometimes appear at odd positions beyond realism which can make existing localization techniques fail to localize. Although this paper is developed for studying caricature face recognition. Our dataset provides 17 landmarks for each caricature, researches on caricature detection and landmark localization can also be done on our dataset.

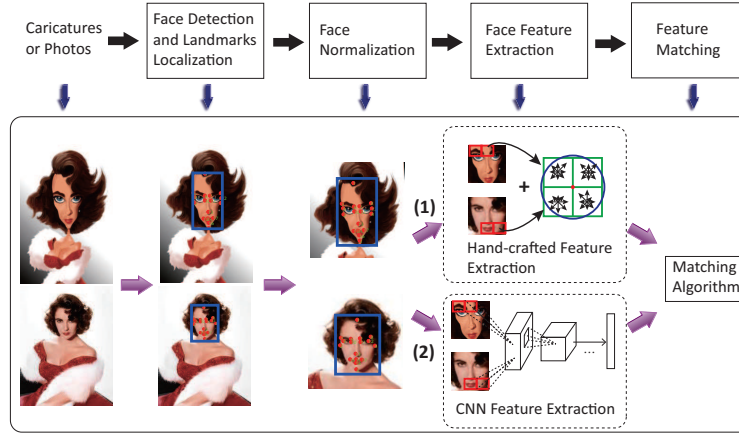


Fig. 3. Caricature face recognition framework used in this paper. First step is face detection and landmark localization. Then faces are aligned. Three alignment methods are tested. With aligned face images, features are extracted. As illustrated, two kinds of feature extraction methods are tested, hand-crafted and convolutional neural network based. Last step is to measure the similarity of extracted features.



Fig. 4. Illustration of misaligned problem. First row, cropped original images. Second row, aligned face images according to eye locations.

4.2. Face Alignment

The second step is face alignment. For this process, the objective is to make the eyes, noses or mouths appear at the same position on two cropped face images, such that the two images are comparable. Three alignment methods were tested in this paper. The alignment process of the first method was to rotate the image first to make two eyes in horizontal positions. After, the image was resized to make eye distance of 75 pixels. Then the cropped region was defined by making the eye center to the upside of the bounding box 70 pixels and the left side of the bounding box 80 pixels. The bounding box is of size 160×200 . This alignment method is denoted as eye location based (short for Eye in the following section). In the second alignment method, the first two steps were the same. Then small regions of 40×40 centered at the labeled 17 landmarks were cropped out. This alignment method is denoted as landmark based (Landmark for short). The third alignment method was done according to the face contour (defined by the face landmarks 1-4). Then this bounding box was enlarged by a scale of 1.2 in both width and height. With the enlarged bounding box, the face image inside the bounding box was resized to 160×200 . This alignment method is denoted as bounding box based (Box for short).

The above alignment methods work well for traditional face recognition. However, in caricature face recognition, as faces are exaggerated. Even after alignment is applied, caricatures can still be misaligned. An illustration of the misaligned caricatures after applying the first alignment method (eye-based) is

given in Fig. 4. As can be seen, photos are fairly well aligned. The eyes of caricatures are also aligned well. However there are mismatches in the nose and mouth areas. Thus for caricature face recognition, more sophisticated alignment methods need to be developed.

4.3. Feature Representation

For face feature extraction, we have tested two kinds of methods. The first is hand-crafted feature extraction. The second is convolutional neural network (CNN) based. For hand-crafted features, four kinds of features were tested, including gray, local binary pattern (LBP) [16], Gabor [17] and SIFT [18]. For CNN, we used the trained model VGG-Face [19]. In our experiments, the feature extraction process was the same for caricatures and photos.

Currently, all the tested feature extraction methods are not cross-modal based. For caricature face recognition, it is important to learn or to find features that are invariant to modality changes.

4.4. Matching Algorithm

The last step is to compute the similarity of two sets of features to decide whether they are from the same person. As the features to be matched are highly influenced by modality changes, removing modality variations presents a challenge for developing matching algorithms for this application. For this step, nine subspace and metric learning methods were tested, including both single modal and cross-modal based methods.

5. Evaluation Protocols

Generally, face recognition can be categorized into two categories: face verification and face identification. For face verification, the objective is to judge whether two face images are from the same person. For face identification, the task is to find the identity of a face image from a set of face images. To promote the study of both caricature face verification and identification, four experimental protocols covering both scenarios are developed.

5.1. Caricature Face Verification

The task of caricature face verification is to verify whether a caricature and a photo are from the same person. Therefore, the algorithm is presented with a pair of images and the output is yes or no. To evaluate the performance of caricature face verification, we have built up two settings which are similar to the settings of LFW [20]. One is image restricted setting and the other is image unrestricted setting. While building up these two settings, the training and testing data are made separated.

In image restricted setting, there are two views. View 1 is for parameter tuning and View 2 is for testing. Pairs of images are provided with the information on whether the pairs of images are from the same person. There is no extra identity information. For training, only the provided pairs should be used. In image unrestricted setting, there are also two views. In this setting, training images together with person identities are given. Therefore, as much as pairs can be formulated for training.

For view 1 of both restricted and unrestricted settings, the proportion of subjects in training and testing were set to near 9:1. For view 2, 10 fold cross validation is used. Therefore, to report results on view 2, training has to be done for ten times. Each time, 9 folds of the data should be used for training and the remaining fold for testing.

For these two settings, researchers are encouraged to use the receiver operating characteristic (ROC) curve, $VR@FPR=0.1\%$ and $VR@FPR=1\%$ to report performances, where $VR@FPR=0.1\%$ corresponds to verification rate (VR) with false positive rate (FPR) equal to 0.1%, and $VR@FPR=1\%$ denotes VR with FPR of 1%.

5.2. Caricature Face Identification

The task of caricature face identification can be formulated into two settings. One is to find the corresponding photos of a given caricature from a photo gallery (denoted as Caricature to Photo or C2P). The second is to find the corresponding caricatures of a given photo from a set of caricatures (denoted as Photo to Caricature or P2C).

For each of the two settings, there are two views. View 1 is developed for parameter tuning and view 2 for testing. While generating data for these two views, subjects were evenly split to training and testing sets. For view 2, the dataset is randomly split for ten times. Thus for results reporting, the algorithm should run for ten times and report the average results. Besides, for C2P, photos are used as gallery and caricatures are used as probe. For each subject in the gallery set, only one photo is selected. Reversely, for P2C, caricatures are used as gallery and each subject has only one caricature in the gallery.

For these two settings, we encourage researchers to use the cumulative match curve (CMC) for evaluation and report average Rank-1 and Rank-10 results.

6. Baseline Performances

Following the established evaluation protocols and the caricature face recognition framework, we evaluated three alignment methods, five feature extraction methods and nine subspace and metric learning methods. Results are summarized into two parts, verification results and identification results.

Table 2. Results of various alignment methods under restricted and unrestricted settings.

Restricted			
Method	FAR=0.1%	FAR=1%	AUC
Gray-Eye	0.0058	0.0478	0.6198
Gray-Landmark	0.0210	0.0745	0.7124
Gray-Box	0.0104	0.0465	0.6626
SIFT-Eye	0.0278	0.1046	0.7383
SIFT-Landmark	0.0467	0.1539	0.7766
SIFT-Box	0.0296	0.0915	0.7195
VGG-Eye	0.2142	0.4028	0.8957
VGG-Box	0.2842	0.5553	0.9463
UnRestricted			
Method	FAR=0.1%	FAR=1%	AUC
Gray-Eye	0.0085	0.0436	0.6233
Gray-Landmark	0.0202	0.0830	0.7185
Gray-Box	0.0142	0.0573	0.6636
SIFT-Eye	0.0274	0.1074	0.7336
SIFT-Landmark	0.0443	0.1524	0.7800
SIFT-Box	0.0342	0.1176	0.7242
VGG-Eye	0.1924	0.4088	0.8980
VGG-Box	0.3207	0.5676	0.9459

Table 3. Results of various feature extraction methods under restricted and unrestricted settings.

Restricted			
Method	FAR=0.1%	FAR=1%	AUC
Gray	0.0058	0.0478	0.6198
LBP	0.0033	0.0192	0.6002
Gabor	0.0323	0.0975	0.7158
SIFT	0.0278	0.1046	0.7383
VGG	0.2142	0.4028	0.8957
UnRestricted			
Method	FAR=0.1%	FAR=1%	AUC
Gray	0.0085	0.0436	0.6233
LBP	0.0019	0.0165	0.5970
Gabor	0.0276	0.1036	0.7178
SIFT	0.0274	0.1074	0.7336
VGG	0.1924	0.4088	0.8980

6.1. Verification Results

6.1.1. Influence of Alignment Methods

Results of the three alignment methods discussed in Section 4.2 are firstly provided in Table 2. The three alignment methods were combined with three feature extraction methods. For extracting Gray features, the resulting aligned face images or patches are simply reshaped into vectors. Under the eye location based and bounding box based alignment settings, to extract SIFT features, the 160×200 image is partitioned into patches of 40×40 with a step size of 20. A 32-dimension SIFT feature is extracted in each patch. Under the landmark based alignment setting, SIFT feature is extracted in each patch and all the features are concatenated. To extract CNN features, the model of VGG-Face is used. Note the landmark based alignment method is not applicable with VGG-Face, as it is developed for extracting features of the entire face. While using VGG-Face, the aligned images are resized to 224, all three color

Table 4. Results of various learning methods under restricted and unrestricted settings.

Method	Restricted			Unrestricted		
	FAR=0.1%	FAR=1%	AUC	FAR=0.1%	FAR=1%	AUC
Euclidean	0.0212	0.0828	0.6613	0.0256	0.0840	0.6611
PCA	0.0467	0.1539	0.7766	0.0443	0.1524	0.7800
KDA	-	-	-	0.0662	0.2423	0.8746
KissME	0.0455	0.1215	0.7236	0.0456	0.1466	0.7812
ITML	0.0508	0.1807	0.8407	0.0535	0.1848	0.8276
LMNN	-	-	-	0.0659	0.2137	0.8419
CCA	0.0477	0.1296	0.7753	0.0502	0.1766	0.8118
MvDA	-	-	-	0.0141	0.0829	0.7534
CSR	-	-	-	0.1176	0.3186	0.8868
KCSR	-	-	-	0.1166	0.3200	0.8882

channels are kept.

In Table 2, the resulting feature vectors are combined with the principle component analysis (PCA) [21] for evaluation. From the table, for hand-crafted features, it is obvious that landmark based feature extraction obtains the best performance compared with the other two. The results of eye location based and bounding box based methods are comparable. In some cases, eye location based results are a slightly better, while in other cases, bounding box based methods are better. This illustrates that landmark based method is the best for alleviating misalignment problem. However, for deep learning based feature, the bounding box based alignment method is much better than eye location based method. This is perhaps because that the bounding box based alignment can crop the whole face into the aligned image. The cropped face image of eye location based method sometimes misses some parts of the face such as chin and mouth due to exaggerations in caricatures. The deep learning based methods may have some kind of mechanisms to alleviate this misalignment problem. Thus the results of bounding box based alignment are better.

6.1.2. Evaluation of Different Feature Extraction Scheme

Then we compared various feature extraction methods, including Gray, LBP, Gabor, SIFT and CNN. For all these feature extraction, we used the same eye location based alignment method for fair comparison. Gray, SIFT and CNN feature extractions are the same as the procedures in the previous section. For LBP feature extraction, uniform LBP was used and LBP was extracted in patches of 40×40 with a step size of 8 pixels. Gabor feature was extracted by resizing the face image to 128×128 and 40 Gabor filters were used, of 8 orientations and 5 scales. After filtering the images, all the filtered image were down sampled to its $\frac{1}{8}$ of its original size. All the extracted features were then combined with PCA for dimension reduction.

From Table 3, SIFT feature is almost the best among hand-crafted features. Gabor is the next. The results of LBP and Gray are similar. Results of VGG-Face is the best and there is a significant improvement over the SIFT. Note that VGG-Face was not fine-tuned for this application. Still the performance is much better than that of any hand-crafted features. However, the results of VGG-Face for FAR=0.1% and FAR=1% are respectively 0.2142 and 0.4028 for image restricted setting and 0.1924 and 0.4088 for image unrestricted setting. For these two

Table 5. Summary of results under restricted and unrestricted settings.

Restricted			
Method	FAR=0.1%	FAR=1%	AUC
SIFT-Landmark-ITML	0.0508	0.1807	0.8407
VGG-Eye-PCA	0.2142	0.4028	0.8957
VGG-Eye-ITML	0.1897	0.4172	0.9114
VGG-Box-PCA	0.2842	0.5553	0.9463
VGG-Box-ITML	0.3494	0.5722	0.9541
Unrestricted			
Method	FAR=0.1%	FAR=1%	AUC
SIFT-Landmark-KCSR	0.1166	0.3200	0.8882
VGG-Eye-PCA	0.1924	0.4088	0.8980
VGG-Eye-KCSR	0.2346	0.4857	0.9246
VGG-Box-PCA	0.3207	0.5676	0.9459
VGG-Box-KCSR	0.3909	0.6582	0.9629

evaluation protocols, there is still rooms for improvement even with deep learning.

6.1.3. Evaluation of Different Learning Methods

In this section, we tested several single view subspace learning methods including PCA [21], kernel discriminant analysis (KDA) [22] and multi-view subspace learning methods such as, canonical correlation analysis (CCA) [23], multi-view discriminant analysis (MvDA) [24], coupled spectral regression (CSR) [14], kernel coupled spectral regression (KCSR) [14]. State-of-the-art single view metric learning methods including KISSME [25], information-theoretic metric learning (ITML) [26] and large margin nearest neighbor (LMNN) [27].

For these experiments, as verified in the previous section, landmark based alignment and SIFT were used in this configuration for comparing all the subspace learning methods. Before applying these methods, PCA was also applied. Results are summarized in Table 4. As the restricted setting only provides with that two images are either of same class or not and some algorithms such as KDA, LMNN, MvDA, CSR, KCSR requires explicit label information for each image, these methods are not applicable for the image restricted setting. From Table 4 the restricted setting part, the best result was achieved by ITML. For unrestricted setting, the best and second best results were achieved by CSR and KCSR. In summary, the results after applying learning methods are all better than that of Eu-

Table 6. Results of various alignment methods under C2P and P2C settings.

C2P		
Method	Rank-1 (%)	Rank-10 (%)
Gray-Eye	4.84 ± 0.56	20.65 ± 1.04
Gray-Landmark	9.99 ± 0.72	33.28 ± 1.24
Gray-Box	5.14 ± 0.56	22.86 ± 1.23
SIFT-Eye	10.36 ± 1.20	35.13 ± 1.55
SIFT-Landmark	15.63 ± 0.82	43.48 ± 1.69
SIFT-Box	10.33 ± 1.05	34.12 ± 1.68
VGG-Eye	35.07 ± 1.84	71.64 ± 1.32
VGG-Box	49.89 ± 1.97	84.21 ± 1.08
P2C		
Method	Rank-1 (%)	Rank-10 (%)
Gray-Eye	4.49 ± 0.49	20.30 ± 1.17
Gray-Landmark	7.96 ± 0.76	30.92 ± 1.85
Gray-Box	5.17 ± 0.71	23.82 ± 1.39
SIFT-Eye	9.06 ± 0.85	32.56 ± 1.27
SIFT-Landmark	12.47 ± 1.14	40.13 ± 1.53
SIFT-Box	8.92 ± 1.11	32.05 ± 1.90
VGG-Eye	36.18 ± 3.24	68.95 ± 3.25
VGG-Box	50.59 ± 2.37	82.15 ± 1.31

clidean which simply uses original features. For unrestricted setting, as the performances of CSR and KCSR are the best as they are designed for cross-view subspace learning. This suggest that further studies on cross-view metric learning or cross-view subspace learning for restricted setting will be beneficial, as there is little work on this topic.

6.1.4. Summary of Results

A summary of the better combinations of the methods at three stages are given in Table 9. The results of deep learning feature based methods outperform the hand-crafted features based method to a large extent. Another observation is that the results of VGG-Face features can be further improved with KCSR. This is mainly because VGG-Face is trained using only photos and may not be able to deal with modality variations. With the help of KCSR to remove modality variations, the performance can be improved. The last is that although deep learning based methods achieve good results. The results of FAR=0.1% and FAR=1% are still far from satisfactory, indicating that there is still space for improvement.

6.2. Identification Results

6.2.1. Influence of Alignment Methods

Results of alignment methods under the identification setting are summarized in Table 6. The evaluation performance measures used include Rank-1 and Rank-10. Under the two identification based settings, for hand-crafted features, landmark based alignment is the best. The results of eye location and bounding box based alignment methods are comparable with each other. For the deep learning based features, bounding box based alignment method is the best.

6.2.2. Evaluation of Different Feature Extraction Scheme

Under caricature to photo and photo to caricature identification settings, four hand-crafted and one CNN based feature

Table 7. Results of various feature extraction methods under C2P and P2C settings.

C2P		
Method	Rank-1 (%)	Rank-10 (%)
Gray	4.84 ± 0.56	20.65 ± 1.04
LBP	4.31 ± 0.72	20.12 ± 1.85
Gabor	8.90 ± 0.78	32.15 ± 1.34
SIFT	10.36 ± 1.20	35.13 ± 1.55
VGG	35.07 ± 1.84	71.64 ± 1.32
P2C		
Method	Rank-1 (%)	Rank-10 (%)
Gray	4.49 ± 0.49	20.30 ± 1.17
LBP	4.31 ± 0.64	19.48 ± 0.98
Gabor	8.26 ± 0.87	30.43 ± 1.77
SIFT	9.06 ± 0.85	32.56 ± 1.27
VGG	36.18 ± 3.24	68.95 ± 3.25

extraction methods are also compared. The alignment method used for this experiments was eye location based. Results in Table 7 show that SIFT is the best among the four hand-crafted features. VGG-Face is much better than SIFT.

6.2.3. Evaluation of Different Learning Method

Influence of different subspace learning methods were also tested. All the nine methods tested in Section 6.1.3 were applied here with the same features. From Table 8, CSR and KCSR achieved similarly the best results. The best rank-1 performance for C2P setting is only 25.18 ± 1.39 and 23.42 ± 1.57 for P2C setting. This means that there is a huge space for improvement for these two settings with traditional face recognition process. Further studies on caricature face recognition will certainly provide findings that can also help reveal more intrinsic clues for traditional face recognition.

6.2.4. Summary of Results

For C2P and P2C settings, we also summarized the results in Table 9. VGG-Face combined with bounding box based alignment method is also better than hand-crafted SIFT feature. The performance of VGG-Face can be further improved with KCSR. Thus, one future direction is to develop deep learning methods that are able to remove modality variations.

7. Conclusions

A caricature dataset of 252 person with 6024 caricatures and 5974 photos is collected. Together, facial landmarks, evaluation protocols and baseline performances are provided on the dataset. Specifically, four evaluation protocols are built. Following these settings, a set of face alignment methods, hand-crafted and deep learning features, and various subspace and metric learning methods are tested. Among all the methods, the best is the bounding box based alignment method combined with VGG-Face feature extraction. The performance of this feature can be further improved with metric learning and subspace learning methods. With this dataset and from the baseline evaluations, there are several future directions. Caricature face

Table 8. Results of various learning methods under C2P and P2C settings.

Method	C2P		P2C	
	Rank-1 (%)	Rank-10 (%)	Rank-1 (%)	Rank-10 (%)
Euclidean	13.38 \pm 1.10	38.40 \pm 1.92	9.04 \pm 0.80	29.63 \pm 1.35
PCA	15.63 \pm 0.82	43.48 \pm 1.69	12.47 \pm 1.14	40.13 \pm 1.53
KDA	19.32 \pm 1.36	56.77 \pm 1.58	18.92 \pm 1.35	57.19 \pm 2.61
KISSME	15.16 \pm 1.63	43.95 \pm 2.13	13.30 \pm 1.18	43.63 \pm 1.64
ITML	15.25 \pm 3.07	46.39 \pm 6.46	16.48 \pm 1.77	49.88 \pm 2.29
LMNN	17.92 \pm 0.86	50.58 \pm 1.72	15.90 \pm 1.73	48.08 \pm 1.95
CCA	10.84 \pm 0.78	40.76 \pm 1.08	10.73 \pm 0.94	41.12 \pm 1.87
MvDA	4.77 \pm 0.74	27.73 \pm 1.90	4.71 \pm 0.87	27.19 \pm 2.55
CSR	25.18 \pm 1.39	60.95 \pm 1.20	23.36 \pm 1.47	60.27 \pm 1.97
KCSR	24.87 \pm 1.50	61.57 \pm 1.37	23.42 \pm 1.57	60.95 \pm 2.34

Table 9. Summary of results under C2P and P2C settings.

C2P		
Method	Rank-1 (%)	Rank-10 (%)
SIFT-Landmark-KCSR	24.87 \pm 1.50	61.57 \pm 1.37
VGG-Eye-PCA	35.07 \pm 1.84	71.64 \pm 1.32
VGG-Eye-KCSR	39.76 \pm 1.60	75.38 \pm 1.34
VGG-Box-PCA	49.89 \pm 1.97	84.21 \pm 1.08
VGG-Box-KCSR	55.41 \pm 1.41	87.00 \pm 0.92
P2C		
Method	Rank-1	Rank-10
SIFT-Landmark-KCSR	23.42 \pm 1.57	60.95 \pm 2.34
VGG-Eye-PCA	36.18 \pm 3.24	68.95 \pm 3.25
VGG-Eye-KCSR	40.67 \pm 3.61	75.77 \pm 2.63
VGG-Box-PCA	50.59 \pm 2.37	82.15 \pm 1.31
VGG-Box-KCSR	55.53 \pm 2.17	86.86 \pm 1.42

landmark detection is of great interest and is a key step for caricature recognition. As the performance on this dataset is still far from saturated, future work and research on caricature face feature extraction and cross-modal metric learning methods are also promising directions.

Acknowledgement

The work was supported by the National Science Foundation of China (Nos. 61432008, 61305068, 61321491) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

- [1] Sun, Y., Chen, Y., Wang, X., Tang, X.. Deep learning face representation by joint identification-verification. In: Proc. Adv. Neural Inf. Process. Syst. 2014, p. 1988–1996.
- [2] Mauro, R., Kubovy, M.. Caricature and face recognition. *Memory & Cognition* 1992;20(4):433–440.
- [3] Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proc IEEE* 2006;94(11):1948–1962.
- [4] Rodriguez, J., Gutierrez-Osuna, R.. Reverse caricatures effects on three-dimensional facial reconstructions. *Image Vis Comput* 2011;29(5):329–334.
- [5] Klare, B.F., Bucak, S.S., Jain, A.K., Akgul, T.. Towards automated caricature recognition. In: Proc. Int. Conf. Biometrics. 2012, p. 139–146.
- [6] Abaci, B., Akgul, T.. Matching caricatures to photographs. *Signal, Image and Video Processing* 2015;9(1):295–303.
- [7] Ouyang, S., Hospedales, T., Song, Y.Z., Li, X.. Cross-modal face matching: beyond viewed sketches. In: Proc. Asian Conf. Comput. Vis. 2014, p. 210–225.
- [8] Crowley, E.J., Parkhi, O.M., Zisserman, A.. Face painting: querying art with photos. In: Proc. Brit. Mach. Vis. Conf. 2015, p. 65.1–65.13.
- [9] Wang, X., Tang, X.. Face photo-sketch synthesis and recognition. *IEEE Trans Pattern Anal Mach Intell* 2009;31(11):1955–1967.
- [10] Wang, S., Zhang, L., Liang, Y., Pan, Q.. Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2012, p. 2216–2223.
- [11] Jin, Y., Lu, J., Ruan, Q.. Coupled discriminative feature learning for heterogeneous face recognition. *IEEE Trans Inf Forensics Security* 2015;10(3):640–652.
- [12] Yi, D., Lei, Z., Li, S.Z.. Shared representation learning for heterogeneous face recognition. In: Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops; vol. 1. 2015, p. 1–7.
- [13] Lin, D., Tang, X.. Inter-modality face recognition. In: Proc. Eur. Conf. Comput. Vis. 2006, p. 13–26.
- [14] Lei, Z., Li, S.Z.. Coupled spectral regression for matching heterogeneous faces. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2009, p. 1123–1128.
- [15] MegviiInc., . Face++ research toolkit. www.faceplusplus.com; 2013.
- [16] Ahonen, T., Hadid, A., Pietikainen, M.. Face description with local binary patterns: Application to face recognition. *IEEE Trans Pattern Anal Mach Intell* 2006;28(12):2037–2041.
- [17] Liu, C., Wechsler, H.. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans Image Process* 2002;11(4):467–476.
- [18] Lowe, D.G.. Distinctive image features from scale-invariant keypoints. *Int'l J Comput Vis* 2004;60(2):91–110.
- [19] Parkhi, O.M., Vedaldi, A., Zisserman, A.. Deep face recognition. In: Proc. Brit. Mach. Vis. Conf. 2015,.
- [20] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49; University of Massachusetts, Amherst; 2007.
- [21] Kim, K.. Face recognition using principle component analysis. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1996, p. 586–591.
- [22] Cai, D., He, X., Han, J.. Speed up kernel discriminant analysis. *VLDB J* 2011;20(1):21–33.
- [23] Hotelling, H.. Relations between two sets of variates. *Biometrika* 1936;321–377.
- [24] Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.. Multi-view discriminant analysis. *IEEE Trans Pattern Anal Mach Intell* 2016;38(1):188–194.
- [25] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.. Large scale metric learning from equivalence constraints. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2012, p. 2288–2295.
- [26] Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I.S.. Information-theoretic metric learning. In: Proc. Int. Conf. Mach Learn. 2007, p. 209–216.
- [27] Weinberger, K.Q., Saul, L.K.. Distance metric learning for large margin nearest neighbor classification. *J Mach Learn Res* 2009;10(Feb):207–244.